big data and
cognitive computing

MDPI

Article

Wine Ontology Influence in a Recommendation System

Luís Oliveira [1,*], Rodrigo Rocha Silva [2,3,*] and Jorge Bernardino [1,3,*]

[1] Polytechnic of Coimbra, Coimbra Institute of Engineering (ISEC), 3030-190 Coimbra, Portugal
[2] FATEC Mogi das Cruzes, São Paulo Technological College, Mogi das Cruzes 08773-600, Brazil
[3] Centre of Informatics and Systems of University of Coimbra (CISUC), 3030-290 Coimbra, Portugal
[*] Correspondence: a21270604@isec.pt (L.O.); rodrigo.rsilva@fatec.sp.gov.br (R.R.S.); jorge@isec.pt (J.B.)

Abstract: Wine is the second most popular alcoholic drink in the world behind beer. With the rise of e-commerce, recommendation systems have become a very important factor in the success of business. Recommendation systems analyze metadata to predict if, for example, a user will recommend a product. The metadata consist mostly of former reviews or web traffic from the same user. For this reason, we investigate what would happen if the information analyzed by a recommendation system was insufficient. In this paper, we explore the effects of a new wine ontology in a recommendation system. We created our own wine ontology and then made two sets of tests for each dataset. In both sets of tests, we applied four machine learning clustering algorithms that had the objective of predicting if a user recommends a wine product. The only difference between each set of tests is the attributes contained in the dataset. In the first set of tests, the datasets were influenced by the ontology, and in the second set, the only information about a wine product is its name. We compared the two test sets' results and observed that there was a significant increase in classification accuracy when using a dataset with the proposed ontology. We demonstrate the general applicability of the methodology to other cases, applying our proposal to an Amazon product review dataset.

Keywords: wine ontology; Weka clustering algorithms; recommendation system; ontology influence; classification via clustering; machine learning

check for updates

Citation: Oliveira, L.; Rocha Silva, R.; Bernardino, J. Wine Ontology Influence in a Recommendation System. *Big Data Cogn. Comput.* **2021**, *5*, 16. https://doi.org/10.3390/bdcc5020016

Academic Editor: Min Chen

Received: 16 January 2021
Accepted: 13 April 2021
Published: 15 April 2021

1. Introduction

Wine has been produced for thousands of years, and its earliest recorded history extends back nearly 6000 years ago. Since then, it has become one of the most popular types of alcoholic drinks in the world. As its popularity has grown, more and more regions started to produce wine, each with its own different types of grapes, soils, climates, fermentation processes and other factors that influence the taste of the wine, which has resulted in an uncountable variety of wines in the world. With such a rich and lengthy history, wine has naturally been embedded in countless cultures around the world, and it is even considered an integral part of some religions. Despite being an alcoholic drink that, when consumed in excess, can deteriorate a person's health, when consumed with responsibility and control, it even has some health benefits [1].

With such a variety in types, uses and applications, and with its popularization due to the World Wide Web, the wine market is being slowly shifted to an e-commerce business model. In 2014, a study revealed that global online wine sales reached 5% of all wine sales totaling, 6 billion dollars in revenues [2]. An e-commerce business depends on its website to make revenue, and with thousands of e-commerce businesses in the world selling the same products category, the website has to stand out and offer the user a fluid, helpful and appealing experience [3]. Consequently, the user must be helped and guided on what s/he wants. If the user does not find it or takes too long to discover the product desired, s/he may lose interest in the site and never return. This is where a recommendation system plays a role. A recommendation system analyzes the user data and assists the user in deciding whether to purchase a product based on previous users' purchases, reviews or web traffic. As previously stated, there is an enormous variety of wines, and a large number of them are very similar. In recent years, recommendation systems have been developed based on domain knowledge and problem-solving approaches. An ontology refers to a body of knowledge describing some domain, typically a commonsense knowledge domain, using a representation vocabulary. Ontologies are used to solve classification, annotation and rendering, and to create different interpretations that make knowledge representation more effective [4]. This is the core relationship between an ontology and a recommendation system and our motivation to build a wine ontology.

An ontology is composed of a relevant concept set representing the characteristics of a given application domain, their definitions and relationships between the concepts [5]. Despite several other applications, an ontology is normally used to [6] (i) share a common understanding of the information structure among different actors; (ii) enable domain knowledge reuse; (iii) make assumptions of the domain explicit; (iv) separate domain knowledge from operational knowledge; and (v) analyze domain knowledge.

Ontologies are very common in e-commerce business websites as they help to display products and their features [7]. The more complete the ontology, the better, since it provides the customer a wide variety of filters to narrow stock search results, which fulfill the customer's preferences. For this reason, ontologies are currently a very popular research topic [6]. Recommendation systems are dedicated to predicting the preference that a user would give to an item [8,9]. An ontology can integrate the use of heterogeneous information and guide the recommendation preference [10].

In this paper, we explore the effects of an ontology in a recommendation system. We have done this by applying four machine learning clustering algorithms (Simple K-Means, Expectation Maximization—EM, Make Density-Based Clusterer, and Farthest First) with classification via clustering to a wine dataset with user reviews. We use clustering algorithms because they are one of the most common exploratory data analysis techniques used to get an intuition about the data structure [11]. It can be defined as the task of identifying subgroups in the data such that data points in the same subgroup (cluster) are very similar, while data points in different clusters are very different. Clustering algorithms can automatically recognize the pattern inside the data so as to analyze the collected data without their labels.

In the first test, our ontology was integrated into the dataset, and then the clustering algorithms were applied. The second test was done by applying the algorithms into a dataset that had no connection to our ontology and only had the name of the wine as information about the product. Therefore, the results demonstrate the influence of the ontology in the datasets. We predict that when adding more information about the product (first test), we will have better classification accuracy.

The main contributions of this work are the following:

- A new wine ontology;
- Evaluating the ontology's impact on precision and execution time;
- Studying the influence of a general product ontology in classification via clustering.

This paper is structured as follows. In Section 2, we present a summary of research papers related to the topic of this work. In Section 3, we present a wine ontology proposal. Section 4 describes the setup of the experiment. Section 5 presents the tests with the different algorithms and a discussion of the results. Section 6 describes the generalization of the methodology to different product categories, namely on an Amazon product review dataset. Finally, Section 7 presents the conclusions and proposes future work.

2. Related Work

In this section, we select research papers that focus on wine ontologies and recommendation systems.

Classification is a data mining task that finds a model for representing and distinguishing data classes or concepts [11]. In data mining, it is usual to use ontologies to classify labels with the relations defined in the ontology. Balcan et al. [12] presented an ontology with annotated classification labels. The semantics encoded in the classification task have the potential not only to influence the labeled data in the classification task but also to handle many unlabeled data. An ontology specifies the constraints between the multiple classification tasks. This classification task produces the classification hypothesis with the classifiers that produce the least unlabeled error rate and thus the most classification consistency.

A recommendation system named Athena was proposed by IJntema [13], which provides ontology-based recommendations for a news feed system. It extends the Herme framework [14] used to build a news personalization service, with the help of an ontology to determine the semantic relations between terms and concepts.

In [15], we proposed a lightweight method to categorize products into beer or wine. It is used a keyword-based method that basically split the product title into words and matched them with words inside a list of categories. For example, if a product title matched with a word inside a beer keyword list, that product would be categorized as beer. This algorithm was built in C++ language, and after analyzing a dataset and categorizing each product, it created two datasets, one for the products classified as beer and the other for the products classified as wine. The experiments used a wine dataset generated by that algorithm, which also contained real data of user reviews of wine products. This dataset was used in the experiments with our recommending system.

In the work of Allahyari et al. [16], the authors presented an ontology-based approach for the automatic classification of text documents into a dynamically defined set of interest topics. The approach considers the use of DBpedia-based ontology, where entities and relations among entities are identified from the text document.

In [17], the authors propose a new hybrid approach to a recommendation system, which combines the collaborative filtering simplicity with the efficiency of the ontology-based recommenders. The proposed recommendation system considers not only users with similar preferences to the active user but also obtains knowledge about the user, their neighbors, products and the relationships between them. This increases the number of recommended products from categories from which the active user has not yet purchased. Their hybrid ontology-based approach combines the application of ontology and the KNN algorithm and starts by creating a user profile that contains, among other attributes, the categories and respective products bought by the user. To recommend products for the user, they needed to find other users (neighbors) that bought at least one common product and then select the other products bought by the neighbor that satisfy the criteria set of the user. Those products have the potential to be recommended, and in order to test their approach, they compared it to a collaborative filtering approach. They concluded that their recommendation system increases the number of products recommended that belong to categories unknown to the active user. The products that are recommended also match user preferences more easily when compared to the collaborative filtering version. However, their approach had the drawback of taking more time to apply the KNN algorithm to find the k-nearest products.

The authors of [18] present an ontology for Social Event Classification, named LODSE (Linking Open Description of Social Events). The basic idea of this ontology is to create a data model that allows the properties definition to describe social events and improve events classification.

In [19], the authors propose an evaluation measure for the performance assessment of multi-annotation classification systems incorporating ontology knowledge. A distance-based misclassification cost was extended from the unilabel to the multilabel case and further enriched with ontology information like its hierarchy, an annotation agreement factor, and penalties for ignoring relationships. Despite the differences between this work and our proposal, this paper allows insights that ontology knowledge can bring advantages in the classification process, which is what the clustering algorithms used in this work must do.

These papers presented different approaches to the use of an ontology in classification and recommendation systems. Even though all these papers present their own ontology, our work differentiates itself by observing the effects of integrating an ontology into a dataset in classification via clustering and comparing it to the same dataset without ontology integration. It is also important to state that none of the papers focus on developing a wine ontology.

3. A New Wine Ontology

This section presents the proposed ontology and describes the tools used to create this ontology and the classes and relationships contained in it.

An ontology refers to an engineering artifact, constituted by a specific vocabulary used to describe a certain reality, plus an explicit assumption set regarding the intended meaning of the vocabulary words. In other words, an ontology defines a representational term set that we entitle concepts [20], providing potential terms for describing our knowledge about the domain. It is basically a way to represent and share knowledge about the domain and its concepts.

The knowledge represented by an ontology comprises the named entities, relationships between them, entity classification and the class hierarchy [16], and it can be divided into four components:

- *Individuals*: the basic objects or instances;
- *Classes*: groups, concepts or object types. These are similar to the classes used in object-oriented programming languages;
- *Attributes*: aspects, properties or characteristics that the object may have in common with other objects;
- *Relations*: links between objects or classes.

Previous research on recommendation systems has attested to the positive influence of domain knowledge on recommendation systems. For example, the preprocessing can benefit from domain knowledge that can help filter out the redundant or inconsistent data [21–23].

The main objective for creating a new wine ontology was to have an ontology that was simple enough to apply to the available wine recommendation databases. Classes are the focus of most ontologies. Classes describe concepts in the domain. For example, a class of wines represents all wines. Specific wines are instances of this class. A class can have subclasses that represent concepts that are more specific than the superclass. For example, we can divide the palate that describes the wine into Body, Flavor and Sugar. The proposed ontology is not supposed to be extremely complex or wide-ranging because its task is simply to be easily adapted to existing wine datasets. In order to create this ontology in OWL, we used *Protégé*, a free open-source ontology editor software.

Our wine ontology contains the following eleven classes:

1. *Brand:* the brand of the wine;
2. *Region:* region where the wine was produced;
3. *Color:* color of the wine (for example, red, rosé or white);
4. *Winery:* building or property which produced the wine;
5. *Year:* the year that the wine was produced;
6. *Castes:* grape castes/types contained in the wine (for example: Chardonnay, Pinot Noir, Syrah, etc.);
7. *Terroir:* A French term that describes the external conditions in which the wine was produced. This class is divided into four subclasses:
    - ○ Climate;
    - ○ Soil;
    - ○ Terrain;
    - ○ Tradition.

    There are many more terroir parameters but these four are the most important ones;

8. *Taste:* palate that describes the wine. This class is divided in three subclasses:
   - Body;
   - Flavor;
   - Sugar.
9. *Alcohol by volume:* a standard measure of how much alcohol is contained in the wine, which is expressed as a volume percent;
10. *Minimum Price:* current retail minimum wine price;
11. *Maximum Price:* current retail maximum wine price.

We can consider each class as an answer to a question, for example, the class region answers the question: Where is this wine produced? Since our ontology was designed to study the wine business, there are some classes that are not common in some other wine ontologies like brand, minimum price and maximum price. These classes are very valuable when it comes to commercial wines.

With the proposed classes, we can establish several relationships between objects. For example, two objects having the same color will be beneficial to the algorithms, because the algorithm will notice similarities between two objects much more easily than without an ontology.

Figure 1 presents all the classes and relations of our ontology. The classes are represented as rectangles with a yellow circle inside them, and relations are represented with orange lines that connect the classes involved. By looking at Figure 1, we can observe a relation between the classes Winery and Region, which inform about the region of the winery where the wine was produced.

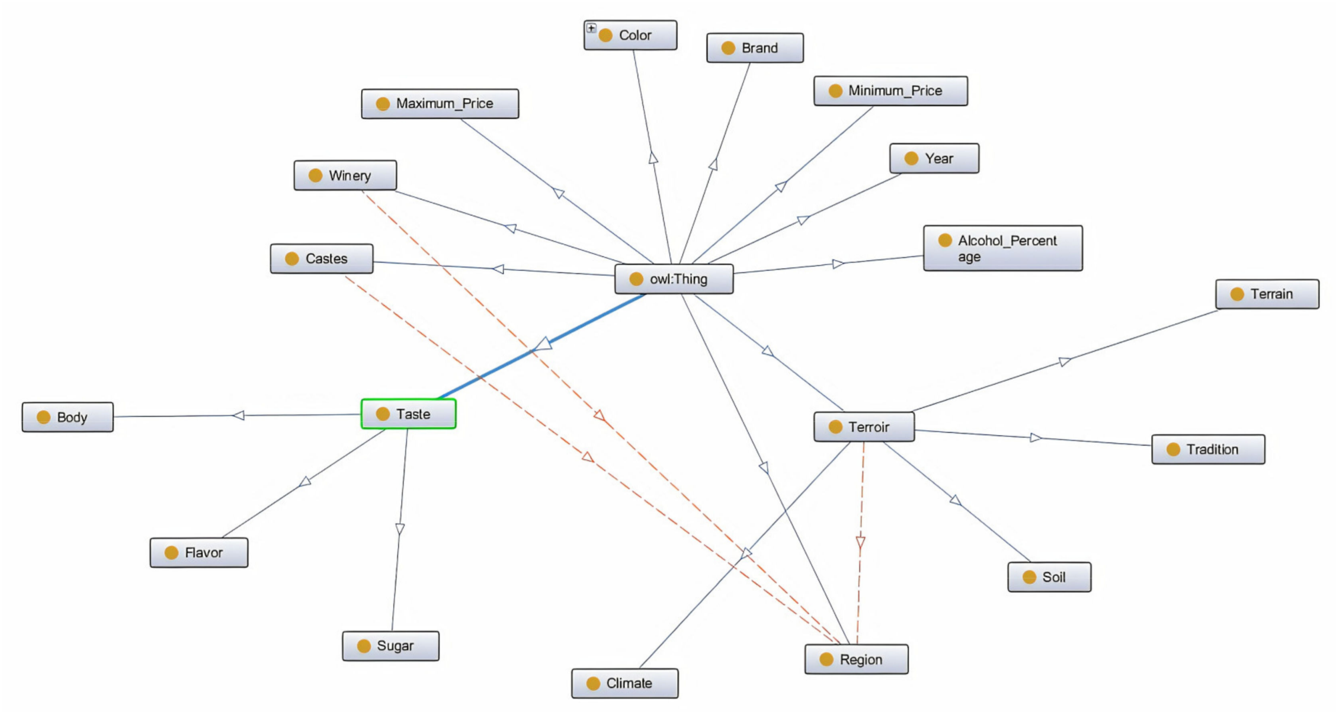


Figure 1. Classes and relations of our wine ontology.

Figure 2 shows the research model of this study matching wine and user's need. The knowledge provided by the expert extracts the ontology proposed. The wine recommendation system using data mining algorithms proposes the wines.

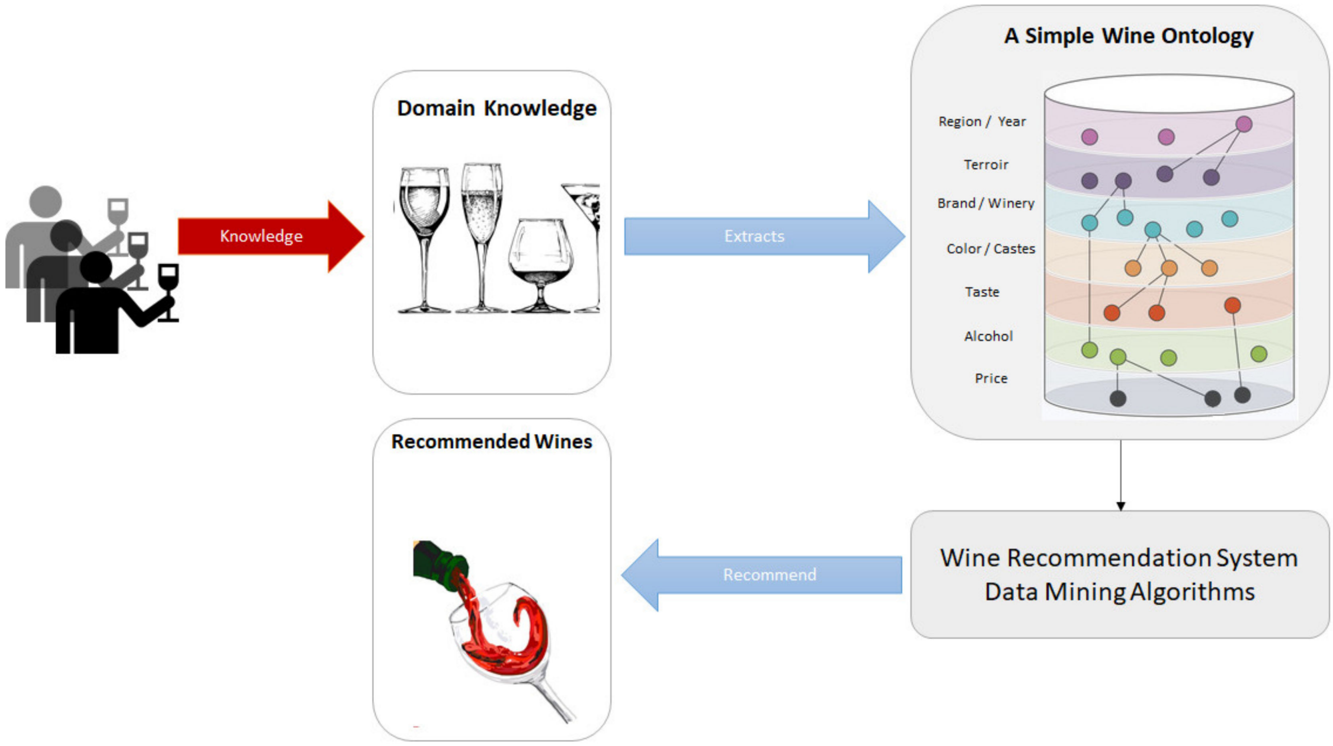


Figure 2. Research model overview of this study.

4. Experimental Setup

This section describes the datasets, hardware, data mining software and algorithms used and the overall experimental setup.

In the experiments, we aim to demonstrate the improvement in classification accuracy by implementing an ontology in a dataset. The experiments consist of applying machine learning algorithms in the datasets, with and without the ontology influence. It is important to state that when we refer to a dataset, we use the term attribute (not to be confused with the attribute concept from the ontology) to describe a column, and when we refer to our wine ontology, we will use the term class. Attribute and class have the same meaning but reference different subjects. Therefore, to clarify, in these experiments, the classes of the ontology are attributes of the data relation.

In order to test the influence of a wine ontology in the application of an algorithm, we will remove attributes from the datasets that correspond to the classes in our ontology (without the ontology influence). We apply the machine learning algorithms and compare the results to the ones from datasets in their original state (with influence). The algorithms have the goal to predict if a user recommends a wine product or not by analyzing the previous reviews contained in the dataset.

4.1. Datasets

We used two datasets in this work, each one with its individual purpose, and both contain user reviews of wine products. The first dataset is based on a combination of the 447_1 and WineReviews datasets, and the second one is a synthetic dataset. As previously stated, the algorithms predict if a user recommends a wine product or not. To do this, the used datasets have six core attributes:

1. *username:* name of the user who wrote the review;
2. *user city:* city where the user lives;
3. *wine name:* name of the wine that is being reviewed;
4. *user rating:* rating the user gave the wine (0 to 5);

5. *did recommend:* denotes if the user did recommend the wine (true or false);
6. *did purchase:* denotes if the user did buy the wine (true or false).

The datasets used for the tests are available online at [24,25].

4.1.1. Real Reviews Dataset

The first dataset is based on a combination of the 447_1 and WineReviews datasets, which can be found in [25]. In [15], we built a categorization system, which analyzed these two small datasets and built a new dataset with only wine products. This dataset with wine products has 941 reviews and 10 attributes. The 10 attributes of the dataset are the six core attributes, mentioned above, and four others that correspond to some classes in our wine ontology, and are the ones that were removed in the experiments. The attributes of this dataset that were removed are the following: brand, color, castes and manufacturer (corresponding to the class winery). This new dataset is referred to in the rest of this paper as the Real Reviews Dataset.

4.1.2. Artificial Reviews Dataset

The second dataset is almost artificial (user reviews are fictitious but the wines are real). The first dataset had real reviews but with a low number of instances. Therefore, we decided to test the clustering algorithms with a large number of instances by building a dataset with random reviews.

We look for a source where we could extract many wines, and we found a Portuguese wine review website called "Blog OsVinhos" [26]. This website has thousands of Portuguese reviews and foreign wines. When we found this website, which had a lot of valuable information, we decided to share this information and contribute to the wine study by building a dataset with the information found on the website. In order to build a dataset, we developed an algorithm that stored all the reviews and its information, from the website. This new dataset was published on a website called "data.world" for public use [24].

Although this dataset had wine reviews from Portuguese reviewers, it had only 2993 instances; at the time, and since we wanted to test the algorithms to a much larger scale than the first dataset, we decided to use only the wines from the "Blog OsVinhos" dataset and build another dataset, which is the one we use in this work, with artificial reviews. We created two lists with random usernames and random cities and then developed a program that randomized reviews, taking the following guidelines into account:

- a username can only have one city;
- a username can only do one review of the same wine;
- the user rating is random;
- there is 50% chance that the user buys the wine;
- if the user reviews a wine with a rating of 5, s/he recommends the wine;
- if the user reviews a wine with a rating lower or equal to 2, s/he does not recommend the wine;
- if the user reviews a wine with a rating of 3 or 4, there is a 75% chance of the user recommending the wine.

The dataset has 1,000,000 reviews and 14 attributes, 6 of them being the core attributes described above. The attributes that are removed in the experiment corresponding to its respective classes of our wine ontology are year, region, producer (corresponds to winery), color, castes, alcohol percentage, minimum price and maximum price. This dataset with one million reviews is referred to in the rest of the paper as Artificial Reviews Dataset.

4.2. The Hardware

In order to execute the experiments, we used a personal computer with the following specifications:

- *RAM:* 16 GB DDR4 SDRAM;
- *CPU:* Intel i7-6700HQ (2.6 GHz, 6 MB);

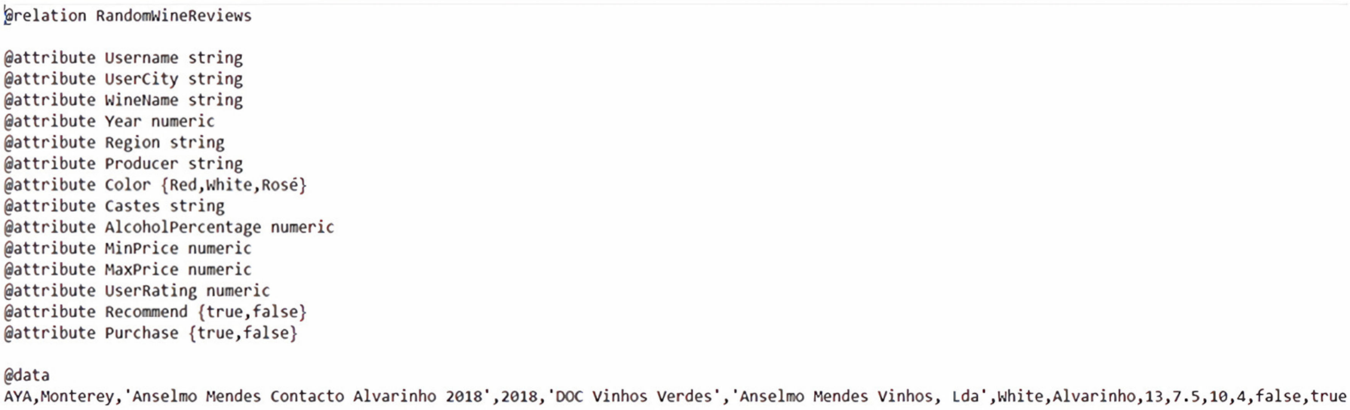
- *GPU:* GeForce GT 950M 2GB.

4.3. The Data Mining Tool

In the tests, we used Weka, which is an open-source data mining software. This software allows running the algorithms on the datasets in an easier way since the Weka GUI (Graphical User Interface) is well organized and provides all the information required to complete the experiments.

To import the datasets into Weka, it is necessary to convert them from CSV (comma-separated values) files into ARFF files, which is a format specific to Weka. The ARFF (attribute-relation file format) format is very similar to the CSV format, which is a common and versatile format used mostly to handle datasets but is more complex since it specifies the data type of each attribute. In Figure 3, we can see an example of an ARFF file. An ARFF file is composed of three segments [27]:

- *Relation:* the first line of the file, which must start with @relation followed by a keyword that identifies the relation or task being studied;
- *Attributes:* a group of lines where each one starts with @attribute followed by the name of the attribute and then followed by the type of the attribute. The attribute can be of the type real (real numbers), nominal (a variety of specific strings), dates or strings;
- *Data:* this is where the instances of the dataset are stored. This segment starts with @data and then the following lines correspond to instances separated by commas with the order specified in the attributes segment.



```
@relation RandomWineReviews

@attribute Username string
@attribute UserCity string
@attribute WineName string
@attribute Year numeric
@attribute Region string
@attribute Producer string
@attribute Color {Red,White,Rosé}
@attribute Castes string
@attribute AlcoholPercentage numeric
@attribute MinPrice numeric
@attribute MaxPrice numeric
@attribute UserRating numeric
@attribute Recommend {true,false}
@attribute Purchase {true,false}

@data
AYA,Monterey,'Anselmo Mendes Contacto Alvarinho 2018',2018,'DOC Vinhos Verdes','Anselmo Mendes Vinhos, Lda',White,Alvarinho,13,7.5,10,4,false,true
```

Figure 3. ARFF file Example.

We used the Weka Java Library to build another program that transforms the datasets that were previously stored as CSV files into ARFF files. Since most of the algorithms in Weka do not support the attribute of string type, we used an unsupervised attribute filter called StringToNominal, which, like its name suggests, converts string attributes into nominal ones by adding each distinct string as a nominal value. In order to remove the attributes in the experiments, we used another unsupervised attribute filter called Remove. Both filters previously mentioned were provided by Weka and can be found in its GUI. The most appropriate parameters were used for all experiments carried out.

4.4. The Algorithms

In the experiments, we used cluster algorithms with classes to cluster evaluation, which can be easily done in the Weka GUI, that try to predict if a user recommends a product or not by finding the most common values to each situation. All the algorithms used were applied with two clusters, one for each value of the recommend attribute (true or false).

We used the following four popular algorithms: Simple K-Means, Expectation-Maximization—EM, Make Density-Based Clusterer, and Farthest First. All these algorithms are briefly explained in the following subsections.

4.4.1. Simple K-Means Algorithm

K-Means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The K-means algorithm identifies k number of centroids and then allocates every data point to the nearest cluster while keeping the centroids as small as possible. Then, it computes a new mean for each cluster. This process iterates until the criterion function converges [28]. The 'means' in the K-Means refers to averaging the data, that is, finding the centroid.

The Simple K-Means Algorithm needs to calculate the distance between the object being analyzed and the cluster mean, and to do so, it uses the Euclidean Distance equation, which is an option present in the Weka GUI. The Euclidean Distance equation is calculated with the following formula:

$$dist((x,y),(a,b)) = \sqrt{(x-a)^2 + (y-b)^2}$$

where (x, y) and (a, b) are the coordinates of two points.

4.4.2. EM Algorithm

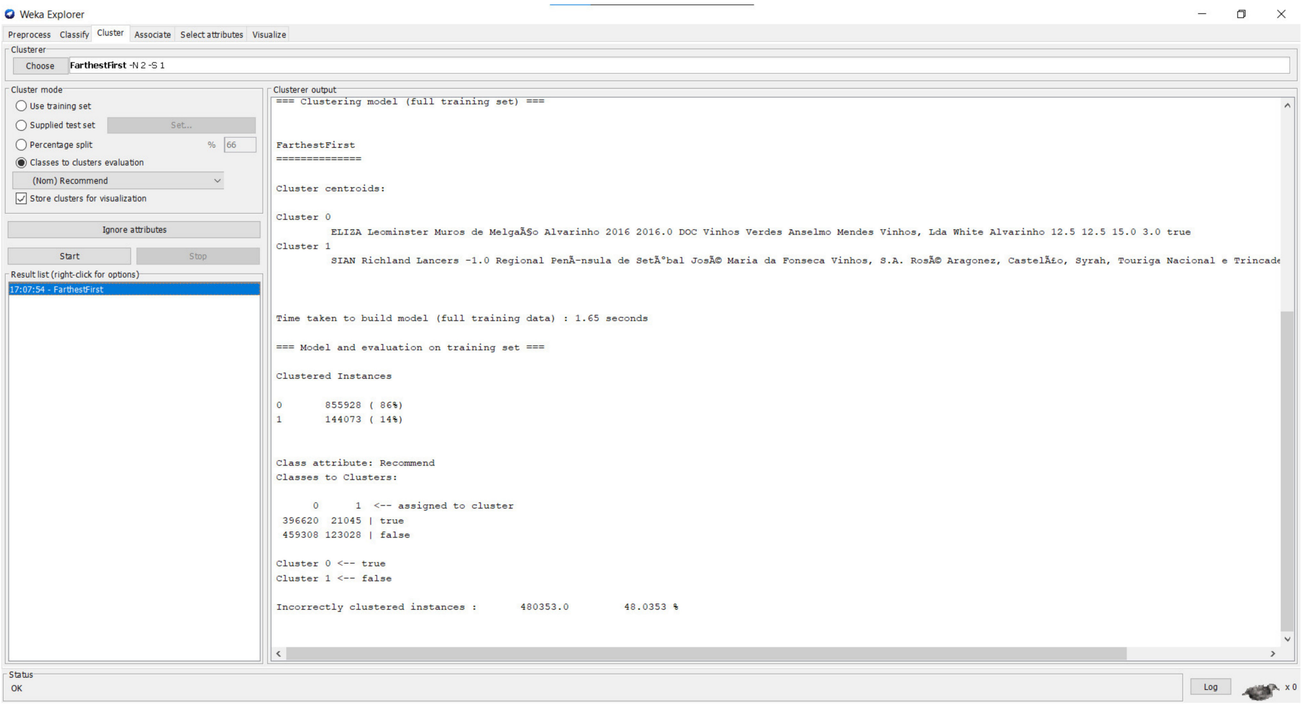
The Expectation-Maximization (EM) algorithm is an interactive method for finding maximum likelihood or maximum a posteriori probability (MAP) that parameters estimate in statistical models, where the model depends on unobserved latent variables. The EM iteration alternates between performing an expectation (E) step, which computes the expectation of the log-likelihood evaluated using the current estimate for the parameters, and maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. These parameter estimates are then used to determine the distribution of the latent variables in the next E step [28].

4.4.3. Make Density-Based Clusterer Algorithm

Make Density-Based Clusterer is a meta-clusterer that wraps a clustering algorithm returning a probability distribution and density. It fits a discrete distribution or a symmetric normal distribution (whose minimum standard deviation is a parameter) for each cluster and attribute [29]. In this case, we wrapped the clustering algorithm Simple K-Means with two clusters.

4.4.4. Farthest First Algorithm

Farthest First is a variant of K-Means that places each cluster center in turn at the point furthest from the existing cluster centers. This point must lie within the data area. This greatly speeds up the clustering in most cases since less reassignment and adjustment is needed. Farthest-point heuristic-based method has the time complexity O(nk), where n is the number of objects in the dataset and k is the number of desired clusters. This method is also fast and suitable for large-scale data mining applications [28]. Figure 4 illustrates the real implementation of our ontology using the Farthest First algorithm in Weka.



Clusterer
FarthestFirst -N 2 -S 1

Cluster mode
○ Use training set
○ Supplied test set
○ Percentage split % 66
● Classes to clusters evaluation
(Nom) Recommend
☑ Store clusters for visualization

Result list (right-click for options)
17:07:54 - FarthestFirst

Clusterer output

```
=== Clustering model (full training set) ===


FarthestFirst
==============


Cluster centroids:

Cluster 0
        ELIZA Leominster Muros de MelgaÃ§o Alvarinho 2016 2016.0 DOC Vinhos Verdes Anselmo Mendes Vinhos, Lda White Alvarinho 12.5 12.5 15.0 3.0 true
Cluster 1
        SIAN Richland Lancers -1.0 Regional PenÃ-nsula de SetÃºbal JosÃ© Maria da Fonseca Vinhos, S.A. RosÃ© Aragonez, CastelÃ£o, Syrah, Touriga Nacional e Trincade



Time taken to build model (full training data) : 1.65 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       855928 ( 86%)
1       144073 ( 14%)


Class attribute: Recommend
Classes to Clusters:

      0      1  <-- assigned to cluster
 396620  21045 | true
 459308 123028 | false

Cluster 0 <-- true
Cluster 1 <-- false

Incorrectly clustered instances :       480353.0        48.0353 %
```

Status
OK

Figure 4. Results of the Farthest First algorithm in Weka.

4.5. Methodology Flowchart

Figure 5 illustrates the methodology used in the experiments. First, we built the ontology and for each dataset applied the Weka filter. Then, we ran the clustering algorithms with and without ontology influence. Finally, we compared the results using precision and execution time metrics.

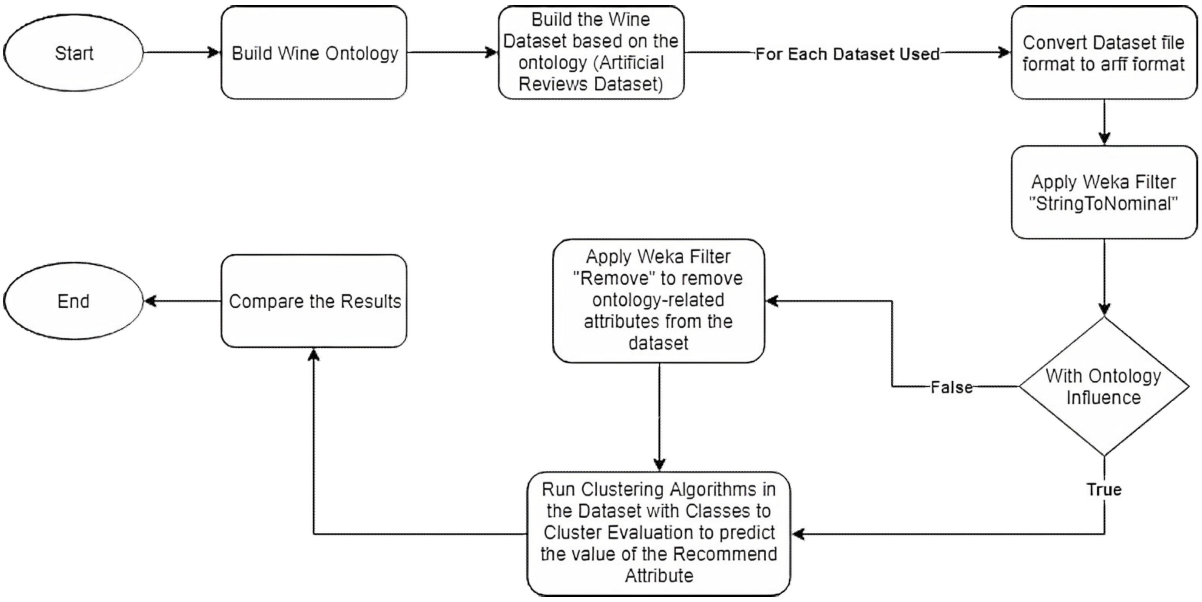

Start
Build Wine Ontology
Build the Wine Dataset based on the ontology (Artificial Reviews Dataset)
For Each Dataset Used
Convert Dataset file format to arff format
Apply Weka Filter "StringToNominal"
With Ontology Influence
False
True
Apply Weka Filter "Remove" to remove ontology-related attributes from the dataset
Run Clustering Algorithms in the Dataset with Classes to Cluster Evaluation to predict the value of the Recommend Attribute
Compare the Results
End

Figure 5. Methodology flowchart.

5. Results and Analysis

This section presents the results of the experiments that were made on the datasets described above. As previously explained, the objective of these experiments was to assess the impact of the ontology on accuracy classification via clustering. We applied the four previously described algorithms on the datasets and presented two metrics: accuracy and execution time. Accuracy in our case is the percentage of incorrectly clustered instances, which is basically the percentage of instances not classified correctly by the algorithms. The execution time is the time to run each algorithm. In the next section, we first describe the experimental results involving the Real Reviews Dataset, followed by the results of the experiments involving Artificial Reviews Dataset.

5.1. Real Reviews Dataset

Table 1 presents the percentage of incorrect clustered instances in the Real Reviews Dataset, with and without ontology influence. The incorrectly clustered instances are the reviews that the algorithm did not predict correctly if the user recommended or did not recommend the wine. The most accurate algorithm with ontology influence was the Simple K-Means algorithm, which had 27.74% of instances incorrectly clustered. This algorithm was also the most affected by the absence of an ontology in the dataset since it has the highest percentage of incorrect clustered instances, namely 39.11%. The Farthest First algorithm seems to have only slightly affected the removal of the ontology since its percentage of error only increased by 0.31%.

Table 1. Percentage of incorrectly clustered instances on the Real Reviews dataset.

Algorithm	With Ontology Influence	Without Ontology Influence
Simple K-Means	27.74%	39.11%
EM	28.80%	36.68%
Make Density-Based Clusterer	27.95%	36.98%
Farthest First	33.48%	33.79%

Table 2 presents the execution time (in seconds) to build the model of each algorithm in the Real Review Dataset, with or without ontology influence. As we can see, since this dataset has less than a thousand instances, the execution time is very low and there is almost no difference between algorithms. It must be noted that Weka can only present this time in seconds and not in milliseconds, and this why some results show 0 s.

Table 2. Time in seconds to build the model on Real Reviews Dataset.

Algorithm	With Ontology Influence	Without Ontology Influence
Simple K-Means	0	0
EM	0.02	0.01
Make Density-Based Clusterer	0.01	0
Farthest First	0.01	0

5.2. Artificial Reviews Dataset

Table 3 presents the percentage of Incorrect Clustered Instances of the Artificial Reviews Dataset for each algorithm, with or without ontology influence. Simple K-Means Algorithm is once again the most precise algorithm with 40.46% incorrectly clustered instances. This time, the less affected algorithm was the Make Density-Based Clusterer with only a 0.35% increase in the incorrectly clustered instances.

Table 3. Percentage of Incorrect Clustered Instances on Artificial Reviews Dataset.

Algorithm	With Ontology Influence	Without Ontology Influence
Simple K-Means	40.46%	49.57%
EM	45.38%	49.95%
Make Density-Based Clusterer	49.60%	49.95%
Farthest First	48.03%	49.89%

Table 4 shows the time taken to build the model of each algorithm in the Artificial Reviews Dataset, with or without the influence of the ontology. In this table, the execution times are varied and much higher than in the other dataset. However, the execution time is still acceptable and fast. The only algorithm that took longer than the others is the EM algorithm, which, with the influence of the ontology, presents 90.95 s to build the model and then had a big drop to 15.26 s when removing the ontology from the dataset.

Table 4. Time in seconds to build the model on Artificial Reviews Dataset.

Algorithm	With Ontology Influence	Without Ontology Influence
Simple K-Means	3.23	1.75
EM	90.95	15.26
Make Density-Based Clusterer	3.97	2.67
Farthest First	1.42	1.10

5.3. Analysis of the Results

Taking into consideration Tables 1 and 3, we can see clearly that the percentage of incorrectly clustered instances is much bigger without the influence of ontology in the datasets. This higher error percentage, without ontology influence in Table 1, is due to the reduced attributes number. Since the attributes number is lower, the algorithms do not have as many reference points to compare cases where users recommend and cases where users do not recommend wines. In other words, with a larger number of attributes, the algorithms can identify more clearly the difference between recommendations and non-recommendations.

Comparing Tables 1 and 3, we can see that the error percentage is much higher in Table 3. Since the Real Review Dataset has only 941 instances, there are not many edge cases (cases that are not similar to the previously analyzed patterns by the algorithms and diverge from the main clusters) compared to the one million instances present in the Artificial Reviews. The fact that the data in the Artificial Reviews Dataset were randomly generated also contributes to a bigger number of edge cases.

When analyzing Tables 2 and 4, which shows the time to build the model for each algorithm, we clearly see that the execution times are smaller without the influence of ontology than with the influence of ontologies. This is caused by the smaller number of attributes in the experiment without the influence of ontology. In the Real Reviews Dataset, we can see that the time to build the model for each algorithm was very close to 0 s, so close in fact that even Weka did not display the milliseconds number in the majority of the experiments. The execution time was obviously higher in the Artificial Reviews Dataset experiments since the algorithms were dealing with a much higher number of instances. Despite having one million instances, Artificial Reviews Dataset experiments did not exceed 4 s to build the algorithm model in most cases. The only case where it did really surpass was with the EM algorithm, which reached 90.95 s in the experiment with the influence of ontology, then reducing to 15.26 s in the experiments without the influence of ontology. The higher execution time is due to the complexity of the EM algorithm, which is an interactive method computing parameters that estimates statistical models. Moreover, the EM iteration alternates between performing an expectation (E) step and a maximization (M) step.

The experiments with the influence of ontology show significant improvements in accuracy classification, using both datasets and with all algorithms. However, the time to build the models does not increase drastically, except when using the EM algorithm (average time increase in Table 4 is 19.70 s and 0.075 s in Table 3).

6. Applying the Methodology to an Amazon Dataset

To prove the general applicability of our methodology to other cases and to disseminate our results, we decided to apply our proposal to a different dataset. Therefore, we can demonstrate that observations that were made in the wine datasets experiments can also be observed in other areas that are not wine-related.

Since our ontology is about wine, it is not appropriated to the dataset that we are going to use, but there are some classes in our wine ontology that correspond to some attributes in the dataset. To test the influence of an ontology in this dataset, we removed the attributes that correspond to classes in our ontology and the attributes that would fit in an ontology related to the dataset.

6.1. Amazon Product Dataset

The new dataset is based on an Amazon Product Review Dataset, which can be found in [30]. We chose this dataset because it has a very similar structure to previous datasets. It contains the same core attributes (wine name attribute corresponds to product name), with the exception of the user city attribute. This dataset has 28,332 instances and 9 attributes, with 5 of them being the core ones previously discussed. The attributes that we removed are brand, categories, primary categories and manufacturer.

6.2. Amazon Dataset Results

Table 5 shows the percentage of Incorrect Clustered Instances of the Amazon Product Reviews Dataset for each algorithm, with or without the influence of ontology. The algorithms have very similar results (all near or equal to 2.83%) when applied with the influence of ontology in the dataset. However, without the influence of ontology, the Farthest First algorithm diverges from the other algorithms results, getting 45.81% incorrectly clustered instances. The experiments with ontology influence show a significant improvement, in accuracy classification terms, for all of the algorithms.

Table 5. Percentage of Incorrectly Clustered Instances on Amazon Product Reviews Dataset.

Algorithm	With Ontology Influence	Without Ontology Influence
Simple K-Means	2.83%	29.87%
EM	2.82%	26.52%
Make Density-Based Clusterer	2.82%	26.86%
Farthest First	2.84%	45.81%

Table 6 shows the execution time taken to build the model of each algorithm in the Amazon Product Reviews Dataset, with or without the influence of ontology. The obtained execution time is not very high even though the dataset has more than 28,000 instances. EM is once again the algorithm with the higher time to build its model, which is expected since it is the most complex algorithm used.

Table 6. Time in seconds to build the model of the Amazon Product Reviews Dataset.

Algorithm	With Ontology Influence	Without Ontology Influence
Simple K-Means	0.23	0.04
EM	0.66	0.32
Make Density-Based Clusterer	0.12	0.05
Farthest First	0.05	0.02

6.3. Analysis of the Results on the Amazon Dataset

By comparing the results, in Table 5, we can see that there is a major improvement in classification accuracy for the experiments with the influence of ontology compared to the ones without it. The biggest increase in the percentage of incorrectly clustered instances was when applying the Farthest First algorithm, which presents a 2.84% classification error with ontology influence, and a large 45.81% classification error without ontology influence. This is a massive 42.97% increase in classification accuracy.

When comparing the execution time shown in Table 6, we can observe the biggest increase in time was 0.34 s, when applying the EM algorithm, which is not significant, despite the fact the influence of ontology has a smaller classification error percentage.

In summary, these results prove that our methodology can be applied in different e-commerce areas with significant improvements in classification accuracy. Therefore, e-commerce companies should invest in ontology use in their business.

When we compare the results obtained with works that use regression algorithms, such as Cortez et al. [31], we observe that the use of classification algorithms together with an ontology makes it possible to obtain accuracy gains within 13% and 36%, with significantly shorter execution time.

7. Conclusions and Future Work

Wine has had a major impact on human civilization. From its primordial production to now and with its commercialization, wine is becoming more complicated with passing time. Ontologies have helped people to distinguish and study different wine types and are now becoming even more valuable with the popularization of e-commerce and recommendation systems. In this work, we show the influence that ontology has on machine learning algorithms, which is the cornerstone of a recommendation system.

The results show that an ontology is critical to increasing the classification accuracy when applying algorithms with classification via clustering. Using the proposed ontology with the Farthest First algorithm, the accuracy classification improved ny 42.97%. However, the execution time to run the classification, in general, was duplicated. This will be studied in the future, but as the wine databases available are small, no more than 5 milliseconds were spent with the Farthest First algorithm.

In conclusion, we demonstrate the importance of e-commerce businesses detailing their ontologies, since a good recommendation system can be the difference between a successful business and a failed business.

As future work, we intend to study how ontology influences memory consumption when applying these algorithms. Weka did not have first- or third-party support to measure the memory usage, and since some algorithms run in milliseconds it was impossible to check the memory usage. We also intend to study more machine learning algorithms.

Author Contributions: Conceptualization, J.B. and R.R.S.; Methodology, L.O. and R.R.S.; Software, L.O.; Validation, L.O., R.R.S. and J.B.; Formal analysis, L.O., R.R.S. and J.B.; Investigation, L.O.; Resources, L.O.; Data curation, L.O.; Writing—original draft preparation, L.O.; Writing—review and editing, L.O., R.R.S. and J.B.; Supervision, R.R.S. and J.B.; Project administration, R.R.S., J.B.; Funding acquisition, J.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://data.world/loliveira1999/portuguese-wine-dataset-from-blogosvinhos [25]; https://data.world/datafiniti/wine-beer-and-liquor-reviews [26]; https://osvinhos.blogspot.com [27].

Acknowledgments: We acknowledge CISUC-Centre of Informatics and Systems of University of Coimbra for the facilities offered during this project.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Soleas, G.J.; Diamandis, E.P.; Goldberg, D.M. Wine as a biological fluid: History, production, and role in disease preven-tion. *J. Clin. Lab. Anal.* **1997**, *11*, 287–313. [CrossRef]
2. Szolnoki, G.; Thach, L.; Kolb, D. Current status of global wine ecommerce and social media. In *Successful Social Media and Ecommerce Strategies in the Wine Industry*; Szolnoki, G., Thach, L., Kolb, D., Eds.; Palgrave Macmillan: New York, NY, USA, 2016; pp. 1–12. [CrossRef]
3. Jennings, M. Theory and models for creating engaging and immersive ecommerce websites. In Proceedings of the 2000 ACM SIGCPR Conference on Computer Personnel Research, New York, NY, USA, 6–8 April 2000; pp. 77–85. [CrossRef]
4. Chen, R.-C.; Huang, Y.-H.; Bau, C.-T.; Chen, S.-M. A recommendation system based on domain ontology and SWRL for anti-diabetic drugs selection. *Expert Syst. Appl.* **2012**, *39*, 3995–4006. [CrossRef]
5. Lim, S.-Y.; Song, M.-H.; Lee, S.-J. The construction of domain ontology and its application to document retrieval. In *Advances in Information Systems*; ADVIS 2004. Lecture Notes in Computer Science; Yakhno, T., Ed.; Springer: Berlin/Heidelberg, Germany, 2004; Volume 3261. [CrossRef]
6. Graça, J.; Mourão, M.; Anunciação, O.; Monteiro, P.; Pinto, H.S.; Loureiro, V. Ontology building process: The wine domain. In Proceedings of the 5th Conference of EFITA, Vila Real, Portugal, 25–28 July 2005; pp. 1138–1145.
7. Noy, N.F.; McGuinness, D.L. *Ontology Development 101: A Guide to Creating Your First Ontology*; Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880; Stanford University: Stanford, CA, USA, 2001. Available online: http://www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness-abstract.html (accessed on 11 November 2019).
8. Adomavicius, G.; Tuzhilin, A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 734–749. [CrossRef]
9. Burke, R. Hybrid Recommender Systems: Survey and Experiments. *User Model. User Adapt. Interact.* **2002**, *12*, 331–370. [CrossRef]
10. Dou, D.; Wang, H.; Liu, H. Semantic data mining: A survey of ontology-based approaches. In Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015), Anaheim, CA, USA, 7–9 February 2015. [CrossRef]
11. Han, J.; Kamber, M.; Pei, J. *Data Mining: Concepts and Techniques*, 3rd ed.; Morgan Kaufmann: Burlington, MA, USA, 2012; ISBN 978-0123814791.
12. Balcan, N.; Blum, A.; Mansour, Y. Exploiting ontology structures and unlabeled data for learning. In Proceedings of the 30th International Conference on Machine Learning, Atlanta, GA, USA, 16–21 June 2013; pp. 1112–1120.
13. Ijntema, W.; Goossen, F.; Frasincar, F.; Hogenboom, F. Ontology-based news recommendation. In Proceedings of the 2010 EDBT/ICDT Workshops (EDBT'10), Lausanne, Switzerland, 22–26 March 2010; Association for Computing Machinery: New York, NY, USA, 2010; pp. 1–6. [CrossRef]
14. Frasincar, F.; Borsje, J.; Levering, L. A semantic web-based approach for building personalized news services. *Int. J. EBusiness Res.* **2009**, *5*, 35–53. [CrossRef]
15. Oliveira, L.; Silva, R.R.; Bernardino, J. Keyword-based Wine and Beer Product Categorization. In Proceedings of the ICMarkTech'20—International Conference on Marketing and Technologies, Lisbon, Portgual, 8–10 October 2020.
16. Allahyari, M.; Kochut, K.J.; Janik, M. Ontology-based text classification into dynamically defined topics. In Proceedings of the 2014 IEEE International Conference, In Semantic Computing (ICSC), Washington, DC, USA, 16–18 June 2014; pp. 273–278.
17. Guia, M.; Silva, R.R.; Bernardino, J. A hybrid ontology-based recommendation system in e-commerce. *Algorithms* **2019**, *12*, 239. [CrossRef]
18. Rodrigues, M.; Silva, R.R.; Bernardino, J. Linking Open Descriptions of Social Events (LODSE): A New Ontology for Social Event Classification. *Information* **2018**, *9*, 164. [CrossRef]
19. Nowak, S.; Lukashevich, H. Multilabel classification evaluation using ontology information. In Proceedings of the ESWC Workshop on Inductive Reasoning and Machine Learning on the Semantic Web, Heraklion, Crete, Greece, 1 June 2009.
20. Ahmed-Ouamer, R.; Hammache, A. Ontology-based information retrieval for e-Learning of computer science. In Proceedings of the 2010 International Conference on Machine and Web Intelligence, Algiers, Algeria, 3–5 October 2010; pp. 250–257. [CrossRef]
21. Khasawneh, N.; Chan, C.-C. Active user-based and ontology-based web log data preprocessing for web usage mining. In Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, Hong Kong, China, 18–22 December 2006; pp. 325–328.
22. Perez-Rey, D.; Anguita, A.; Crespo, J. Ontodataclean: Ontology-based integration and preprocessing of distributed data. In *Lecture Notes in Computer Science*; Maglaveras, N., Chouvarda, I., Koutkias, V., Brause, R., Eds.; Springer: Berlin/Heidelberg, Germany, 2006; Biological and Medical Data Analysis. ISBMDA 2006; Volume 4345. [CrossRef]
23. Amoretti, M.C.; Frixione, M. Representing wine concepts: A hybrid approach. *Appl. Ontol.* **2020**, *15*, 475–491. [CrossRef]
24. Oliveira, L. Portuguese Wine Reviews Dataset from BlogOsVinhos, Data World. Available online: https://data.world/loliveira1999/portuguese-wine-dataset-from-blogosvinhos (accessed on 11 July 2020).
25. Wine Reviews and 447 Datasets. Available online: https://data.world/datafiniti/wine-beer-and-liquor-reviews (accessed on 11 July 2020).
26. Blog OsVinhos. Available online: https://osvinhos.blogspot.com (accessed on 16 July 2020).

27. Santos, R. Weka na Munheca: Um Guia Para uso do Weka em Scripts e Integração Com Aplicações em Java. Available online: https://www.passeidireto.com/arquivo/2389961/weka-na-munheca (accessed on 7 October 2019).
28. Sharma, N.; Bajpai, A.; Litoriya, M.R. Comparison the various clustering algorithms of Weka tools. *Int. J. Emerg. Technol. Adv. Eng.* **2012**, *2*, 73–80. Available online: https://www.researchgate.net/publication/293173843_Comparison_of_the_various_clustering_algorithms_of_weka_tools (accessed on 16 July 2020).
29. Frank, E.; Hall, M.A.; Witten, I.H. *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed.; Elsevier Morgan Kaufmann: San Francisco, CA, USA, 2011.
30. Amazon Reviews Dataset. Available online: https://data.world/datafiniti/consumer-reviews-of-amazon-products (accessed on 11 July 2020).
31. Cortez, P.; Teixeira, J.; Cerdeira, A.; Almeida, F.; Matos, T.; Reis, J. Using Data Mining for Wine Quality Assessment. In *Discovery Science*; Gama, J., Costa, V.S., Jorge, A.M., Brazdil, P.B., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2009; Volume 5808. [CrossRef]